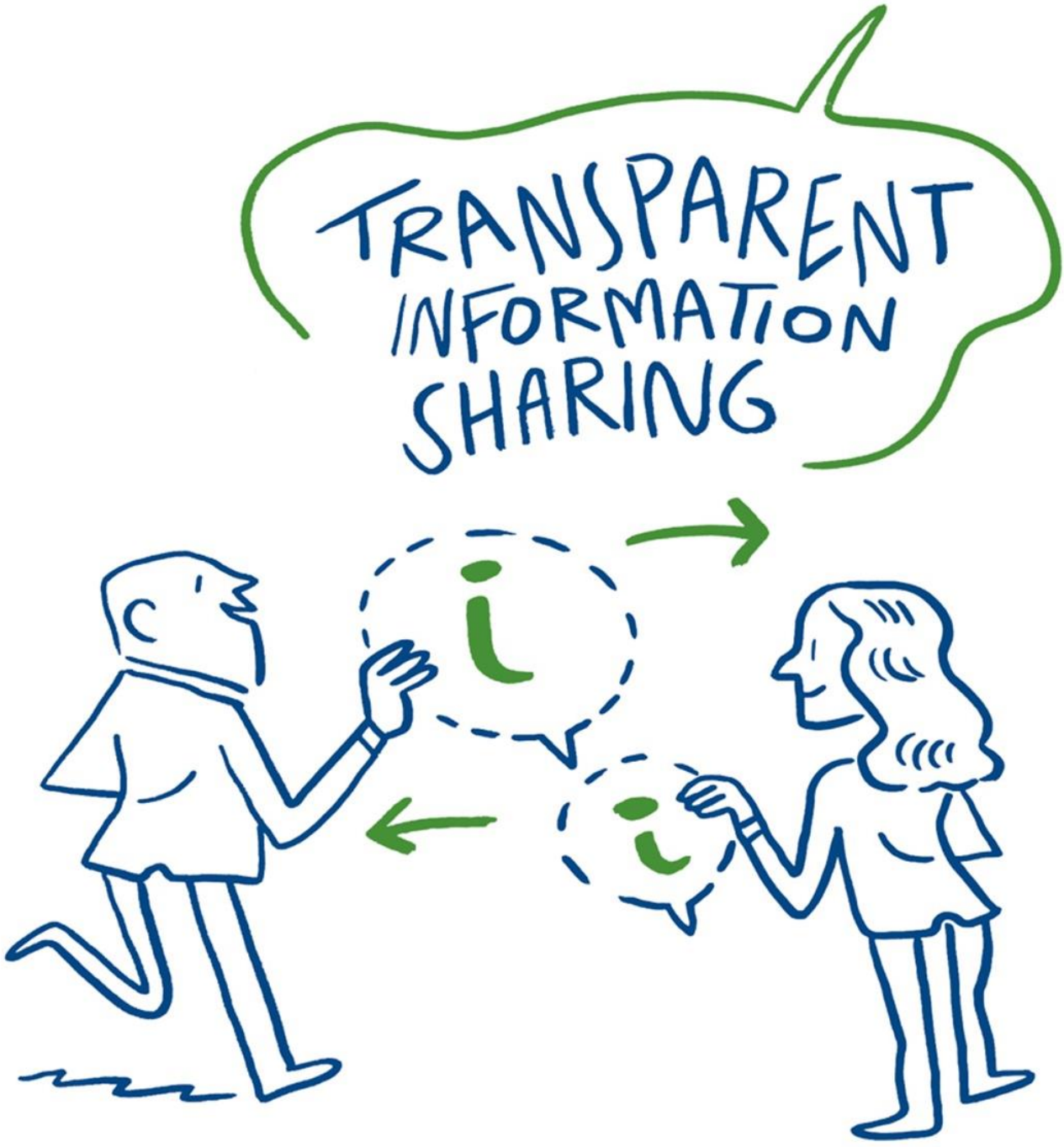
transition
Moving to a smart future



Historic Baselining Methods - Performance Assessment

August 2022

Authors: SSEN & TNEI

tnei

Scottish & Southern Electricity Networks

Contents

Table of Figures

1 Introduction

1.1 Executive Summary

This report, produced in collaboration between SSEN and TNEI, examines the performance of the historic baseline methodologies currently being implemented in the TRANSITION trial periods and those available on the ENA Flexibility Baselining Tool. Both historic and assigned baselines (including both zero and nominated baselines) are available in the ENA Tool, but only the performance of the historic methods is examined here.

The baseline of an asset is an estimate of what an asset's behaviour would have been if it was not providing flexibility. This can then be compared to an asset's metered data to determine the flexible response delivered by the asset during a flexibility event.

As the historic methodologies for baselining are now being used in practice it is important to understand how well they perform – that is, how well they capture the behaviour of a given asset. Critical to this is the error within the calculated baseline – in other words, for periods where flexibility was not being provided, what would the calculated baseline be? The different between the actual measurement and the calculated baseline is the error.

The analysis has shown that, in many cases, the error in the baseline estimate is typically small, centred around zero without significant bias. However, this is not consistent for each asset type and large errors are possible, particularly for PV generation. In most cases the choice of method has a strong effect on consistency and accuracy of results. Overall, the historic methods seem to perform best for hydro and large-scale demand, but results for PV are poor.

The analysis has also shown that the tendency of the methods to over or underestimate the measured data is not consistent across asset types. Applying a Same Day Adjustment (SDA) to the historic method helps to minimise this tendency for most results for some types of assets, but it also introduces severe under and/or over prediction in edge cases.

The results presented in this analysis cover a range of asset types and historic years, however, the sample size for some asset types is very small. Therefore, the results discussed here may not be conclusive for those assets where only a few examples are available e.g., hydro, demand, wind and storage. Further investigation of the historic baseline method performance is recommended, using a wider range of asset data across each of the available asset types.

A further recommendation of this report is the development of external data-based methods for baselining, such as regression-based methods, as the results presented here have shown that there are opportunities to improve upon the performance of the historic methods. Regression-based methods are more complex and data intensive than the simpler historic methods, but could produce more accurate baseline estimates, especially for those assets types such as PV that are not very well served by historic methods. It is important that baselining methods are developed with simplicity, transparency, and replicability in mind, but also inclusivity – and regression-based methods could improve inclusivity, in particular for PV as well as other weather driven renewables or highly variable assets.

1.2 Project Background

The GB network continues to evolve, and there is a clear need for networks to adapt, become more flexible, enhance operations and allow new market models such as peer-to-peer trading to emerge. The 'fit-and-forget' approach of traditional network operation relied on predictable energy use and

production that matched that use. The transition to DSO (Distribution System Operator) model from a DNO (Distribution Network Operator) model has the potential to bring significant benefits to customers; it also brings a range of new complex challenges, unintended consequences and risks for market participants, new entrants and the network licensees.

The ENA Open Networks Project (ON-P) is focussed on defining the DNO transition to a DSO model and has been endorsed by the UK Government's Smart Systems and Flexibility Plan. TRANSITION is designed to help inform the work of the Open Networks Project in the transition, in particular, the project will design and demonstrate the tools and practices DNOs will need to adopt to become DSOs as well as trialling the ON-P market models.

TRANSITION is an Ofgem Electricity Network Innovation Competition (NIC) funded project. Led by SSEN in conjunction with our project partners ENWL, CGI, Origami and Atkins.

In addition, the project is also closely collaborating with the Local Energy Oxfordshire (LEO) project, a UK Industrial Strategy funded project. Both TRANSITION and LEO have objectives that are closely aligned and when combined significantly enhance overall learning. Integration with Project LEO significantly enhances testing opportunity and offers insights into the needs of local energy actors.

The TRANSITION and LEO trials will be used to provide an evidential base on the market dynamics associated with contracted flexibility. In doing so, TRANSITION will build upon previous innovation programmes funded by Ofgem, including New Thames Valley Vision and Low Carbon London, to validate the requirements for DSO systems and management of commercial arrangements for the transaction of flexibility services by multiple market actors.

1.3 Report Overview

TNEI and SSEN have prepared this report examining the performance of the historic baseline methodologies currently being implemented in the TRANSITION trial periods and those available on the ENA Flexibility Baselining Tool.[1] These methodologies and tools have been implemented by TNEI for SSEN throughout 2021 and 2022 as part of the TRANSITION Network Innovation Competition project. The rest of this report is structured as follows:

- Section 2 provides an overview of the context for this work and the motivation for performing this analysis, as well as a summary of the historic baselining methods that have been examined and the data available for the analysis. A brief review of relevant industry reports on baselining performance assessment is included.
- Section 3 describes the methodology used to perform the analysis, defines the metrics used to measure performance, and includes a summary of the datasets used.
- Section 4 presents the results and is divided into three subsections: discussion of overall results, detailed examples of results by different asset type, and finally any other explorative analysis.
- The report closes in Section 5 by outlining key conclusions and recommendations.

[1] ENA Flexibility Baselining Tool

2 Context for This Work

Delivery of Flexibility Services is being explored through TRANSITION, as well as the other related TEF Ofgem NIC (network innovation competition) projects like EFFS and Fusion, and business-as-usual network initiatives such as ENA's Open Networks programme.

Several methodologies have been proposed for determining the baseline of an asset when providing flexibility. The baseline of the asset is an estimate of what an asset's behaviour would have been if it was not providing flexibility. This is required to be able to determine the flexible response delivered by the asset, which is used for settlement and payment.

The focus of this report is a set of baselining methods based on historic data that are being implemented and trialled through TRANSITION. The work in this report examines the performance of those methods, for a range of asset types.

TNEI has been supporting SSEN in considering baselines since late 2020. In addition to carrying out this error analysis, we have produced a Python package for carrying out baseline calculations and implemented this with a tool, hosted online.[1] This was delivered in collaboration with the ENA's Open Networks project in early 2022 as part of the 2021 programme of work in Ena Open Networks Project WS1a Product 7.

2.1 Motivation

Several baseline methods based on historic asset data were recommended for adoption by the ENA in a report carried out for the ON-P project, published in late 2020.[2] These are relatively simple methods with modest data requirements and are therefore easy to implement. As these methods are being implemented in TRANSITION's trial periods, and through ENA's Open Networks programme, it is therefore important to consider how well these methods perform – that is, how well they capture the behaviour of an asset. Critical to this is the *error* within the calculated baseline – in other words, for periods where flexibility was not being provided, what would the calculated baseline be? The different between the actual measurement and the calculated baseline is the error.

Figure 1 Defining the error in the baseline estimate

[2] "Baseline Methodology Assessment Report", DNV-GL, December 2020

2.2 Overview of the Historic Baselining Methods

Following the recommendations in DNV-GL's 2020 report to ENA,[2] the following historic baseline methods have been implemented for flexibility baselining in the TRANSITION project:

- NTVV:[3] A rolling historical baseline which uses data from the last 6 of 10 days, where days are ranked by most similar total energy consumption to the event day. This method applies a same day adjustment (SDA),[4] with a 4-hour window.
- Mid 8-in-10:[5] A rolling historical baseline which uses data from the "middle" of the last 8 of 10 days. Days can be ranked by either peak or average energy, and ranking by peak is the default.
- Mid 8-in-10 with Same Day Adjustment: A rolling historical baseline which uses data from the "middle" of the last 8 of 10 days, but also applies an SDA with a window of 2 hours.
- Mid X-in-Y: A custom rolling historical baseline, where the user can choose how many days to consider and what length of same day adjustment to use.

Each of these methods excludes data from previous days where an asset was providing flexibility from the calculation of the baseline.

It is worth nothing that there are variations in which methods are available in the ENA Flexibility Baselining tool and those available and used by TRANSITION in the Trial Periods. NTVV is not in the ENA Tool. NTVV method was used in TRANSITION's Trial Period 1 but will not be available for Trial Period 2 onwards. The Mid X-in-Y method is available in the ENA Tool for users to explore different combinations of days, with and without applying SDAs.

Two additional methods are also available in the ENA Tool – a nominated baseline and a zero baseline. Both are "assigned" methods, which is an umbrella term to describe approaches where the baseline is assigned without considering the tool considering historic data. The zero method assumes a baseline of zero during the flexibility event. A nominated baseline means that it has been provided/nominated by the FSP, and the baseline could be calculated in a number of different ways. It is possible that an FSP could use regression-based methods, methods that use external data such as temperature and weather data, to build a model from which a baseline can be estimated, but it would be the FSP determining that baseline and not the DNO.

[3] SSEN's New Thames Valley Vision (NTVV) Project implemented a baselining method using historic data, the details of which are described in "Oxfordshire Programme Commercial Arrangements", SSEN, July 2020

[4] A same day adjustment shifts the baseline estimate up or down to match the asset data on the event day, within the specified window. E.g., for NTVV, the SDA ensured that the average demand in the four hours prior to the event matches the average demand at the same time in the 6 selected days.

[5] A detailed description of both Mid 8-in-10 methods is presented in: "Flexibility Baselining Tool – Mathematical Specification", TRANSITION, TNEI and ENA, February 2022

2.3 Examples of Industry Baseline Performance Assessment

There are few industry examples of the performance assessment of baselining methodologies. A brief review of the available sources is outlined below.

ELIA's Baseline methodology assessment consultation report[6] provides a qualitative and market analysis of several available historic baselining methods. This includes a substantial review of the available types of baselining methods beyond those purely based on historical metered data.

Out of the historic methods assessed, it concludes that high X in Y methods are simple but perform well for a number of assets and have the best overall performance. These methods are applied in several flexibility markets internationally for these reasons. The report notes that there is some concern over the accuracy of historic methods for certain asset types, for example those with high variability or dependency on weather conditions. (In particular, it states that an X in Y methodology works well because weather is unlikely to change over the course of a week, or within a shorter window of time – but this is certainly not true in the UK context.) Therefore, there is a gap in the available methods – further methods are needed that remove barriers to participation for assets with high variability. This would include those that use external data – e.g., regression-based methods. The report states that, at the moment, "declarative" methods (known as assigned methods in the ENA Tool, which includes nominated baselines) are seen to be the most inclusive.

Discussion of same day adjustments is presented, with both additive and multiplicative scaling, and the report states that SDAs allow the baseline to be better calibrated to the actual data on the event day, prior to the event. This is true as the SDA represents an up/down shift, or scaling, of the baseline to match the metered day prior to the flexibility event. However, it has the potential to introduce larger errors as no limits are placed on the shift (e.g., generation data could be shifted down to negative values which are not possible in practice). The report considers this but only with respect to multiplicative scaling. It should be noted that the methods outlined in Section 2.2 above include a simple shift as the same day adjustment, and no additional complexity or limits on the SDA are included.

The report concludes that there is limited scope for improving the historic methods themselves as SDAs are already considered. However, it may be possible to introduce some improvements to the use of SDAs, for example by placing limits on the baseline values after scaling. There is also scope for reducing the baseline errors by using other methods.

The ELIA report also offers insight into the practical implications of notification periods and submission/calculation of baselines and the opportunities for manipulation, and provides recommendations on submission and calculations times based on the flexibility product and notification period.

A report by DNV-GL from 2017[7] considers demand loads at peak times and the associated errors in calculated baselines. It concludes that unadjusted historic baselines (those without SDAs) are the least accurate and least precise of the historic methods. The motivation for applying unadjusted baselines would be to ensure that there is no opportunity for manipulation of the baseline through behaviour prior to the event window, but also that behaviour that might be reasonable for some assets (e.g. pre-

[6] "Baseline methodology assessment", ELIA, September 2021

[7] "Evaluation of 2017 Demand Response Demonstrations: C&I Connected Solutions", DNV-GL, February 2018

emptive heating/cooling, or charging prior to delivering flexibility later on) would impact any adjustment applied.

DNV KEMA's Demand Response verification assessment from 2013[8] provides a good qualitative description of what baselines are, why they are needed, as well as why they are bound to have an error in their estimation of an asset's behaviour. The report explains that in order to measure the performance of any demand side resource providing flexibility it is necessary to compare the observed load of the asset to an estimate of the load that would have occurred, in theory, if the asset was not providing flexibility. The estimate of the theoretical load will be subject to error, but this error can be understood and managed. The report goes on to discuss the impacts of SDAs with broadly the same conclusions as the other reports, that they achieve some improvement to baseline as it can be matched to behaviour on the event day, however they note the potential of the SDAs to introduce considerable errors, but this is not supported with quantitative analysis. This report considers baselines applied to demand response only.

The most extensive quantitative analysis was found in a report from KEMA to PJM discussing the Empirical Analysis of Demand Response Baseline Methods[9] that covered 4,500 example datasets. As the title states, this only focuses on demand response. The analysis tries to determine whether they can segment the data so that certain customer types can be said to be better suited for certain methods, but no definitive conclusions are drawn (and no other asset types are considered). The report also states the improved accuracy of using SDAs.

In general, previous analysis shows that baselines will inevitably have errors, but that the magnitude of the error is only one of several important factors that determine which methods should be adopted, including simplicity and transparency, and whether the method introduces any opportunities for manipulation.

[8] "Measurement and verification for Demand Response", DNV KEMA, February 2013

[9] "PJM Empirical Analysis of Demand Response Baseline Methods", KEMA, April 2011

3 Methodology

This section outlines the approach followed for assessing the performance of the historic baseline methods described above.

3.1 Performance Assessment Approach

A baseline is required for flexibility events as it is not possible to simultaneously observe both the actual behaviour of the asset and its "normal" behaviour – we therefore need an estimate of this "normal" behaviour to know how much flexibility has been delivered. This estimate is known as the baseline. The response of the asset is then determined by comparing the baseline with the metered data of the asset during a flexibility event – the difference is then the delivered flexibility.

For this analysis, we use a similar approach to determining the delivered flexibility of the asset. For this analysis, we use a similar approach to determining the delivered flexibility of the asset. First, perform the same calculation of the baseline. The next step would usually be to determine the response from the asset (the flexibility the asset delivered) by comparing the baseline and the metered data during the flexibility event. Instead, in order to determine the error in the baseline, we calculate the baseline and then compare it to the metered data from the day where no flexibility event occurs. This is then the error between the baseline and the asset's actual behaviour.

The aim of the quantitative analysis is to explore the behaviour of the methodologies under perfect conditions where all historical data is available. For each dataset available, each method in question is applied to every single day across the timeseries. For every day of the data, all data is made available except when missing assessment or eligible days are defined.

By considering all of the available historic data in this way, it is possible that baselines may be calculated for days or times when a flexibility event is unlikely, which would then impact the errors calculated. A consequence of this method is that the performance may be performed on days where baselines are simply not appropriate. Efforts have been taken to filter out this data as it obfuscates meaningful and 'realistic' findings. These efforts include removing days where measurement is zero throughout and dropping the 95 percentiles of absolute error for each method.

A second, more explorative qualitative, form of analysis was used to explore the behaviour of methodologies under 'realistic' conditions where the asset successfully provides a given response and may have changed behaviour outside of the event window in order to do this, i.e. shift/spread unused demand backwards/forwards. This is an important consideration for demand or storage assets as flexibility is usually a shift in the consumption of energy, rather than an avoidance, therefore it is important to consider behaviours that may be acceptable in preparation for providing flexibility as well as the potential for gaming.

All analysis was conducted using Python in the form of Jupyter Notebooks. Python is an extremely popular programming language which stands out for its readability, mathematical capabilities, and effective visualisations.

3.2 Performance Metrics

To assess the overall performance of the historic baseline methodologies under specific conditions, the error of each analysis is calculated, processed, and aggregated into appropriate groupings by similar time periods, for example weekdays and weekends, etc. These results can then be examined to identify the overall accuracy of the baseline methods estimation, its variance, any bias that exists, and to identigy any interesting behaviours.

.

3.2.1 Defining the Error

The error of the baseline with respect to the actual measurement value, that is the difference between the baseline estimate and the measured data. This error reflects the polarity of the measured data so that the state of the asset is considered i.e., storage assets where polarity is both positive and negative. For the error, all negative values are considered underpredictions and positive values are overpredictions. This gives an accurate indication of bias in the results.

Error (e) is defined as the difference between the actual measured value (m) of the asset against the estimated baseline value (b), defined as:.

$$e_t = b_t - m_t$$

This can be calculated for each of the timesteps (t), and each of the available historic baseline methods as described in Section. 2.2

3.2.2 Additional Metrics

In addition to the error as defined above, a number of other metrics can be calculated to examine and quantify the baseline error. The descriptions below cover the key metrics presented in Section 4 Analysis and Discussion

Root Mean Square Error (RMSE)

The Root mean square error is calculated for each flexibility event as a measurement of distance of the predicted value from the actual value.

Normalised Root Mean Square Error (nRMSE)

The normalised RMSE values are calculated as the RMSE divided by the average of the measured data for that asset. Normalised values are used so that the results for different types of assets can be compared.

R Squared

R squared values are a measure of how well an estimate is able to capture the variability of the data it is trying to estimate – often referred to as the as measure of the "goodness of fit". A value of 1 indicates the trend is captured fully. A negative score shows a failure to capture the variability.

3.3 Summary of the Data used for this work

SSEN's TRANSITION Team provided several historic data sets for a range of assets across their network, both generation and demand. The historic data provided was for time periods when the assets were not providing flexibility. This was important for the performance assessment so that the baselines calculated in the analysis could be compared to the actual measured data.

Figure 2 below provides a summary of the data used in this analysis, by asset type and number of available datasets.

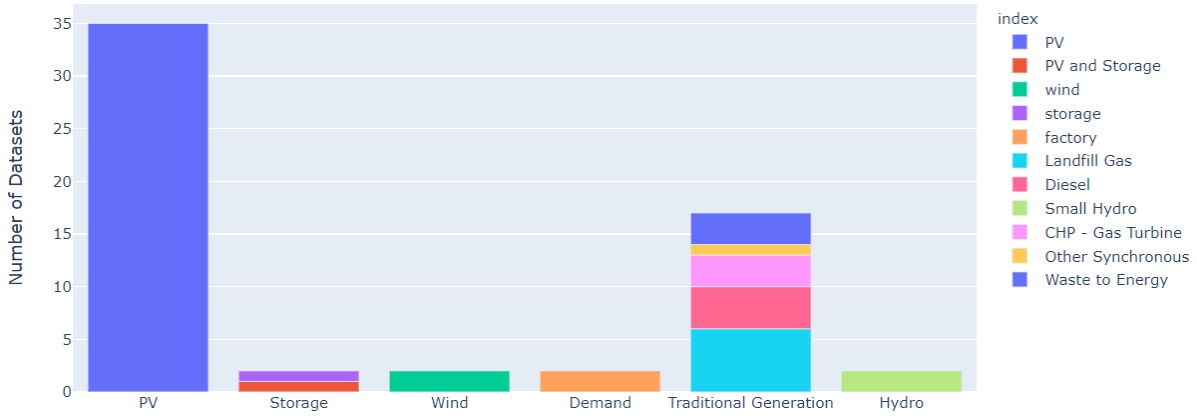


Figure 2 Summary of Data

4 Analysis and Discussion

This section provides an overview of the analysis and key findings, including illustrative graphs of the results.

4.1 Performance of the Historic Methods for each Asset Type

The first aspect of the analysis presented here is the spread of the errors. Applying the Mid 8-in-10 historic method with and without an SDA, the graphs below show box plots[10] of the normalised RMSE, for each asset type.

Normalised values have been plotted so that the errors for each asset type can be compared. The normalised RMSE values are the RMSE divided by the average of the measured data for that asset. Absolute values have been plotted, therefore all errors are above zero. The data used in the plots excludes the few extreme outliers which are produced when normalising the results as this would make the plots unreadable[11].

Figure 3 shows the results for all asset types, except PV which is shown in Figure 3. The plots show mixed impact when applying an SDA. The SDA can reduce the magnitude of the error, as we can see in the case of Hydro in particular, however this varies significantly by asset type.

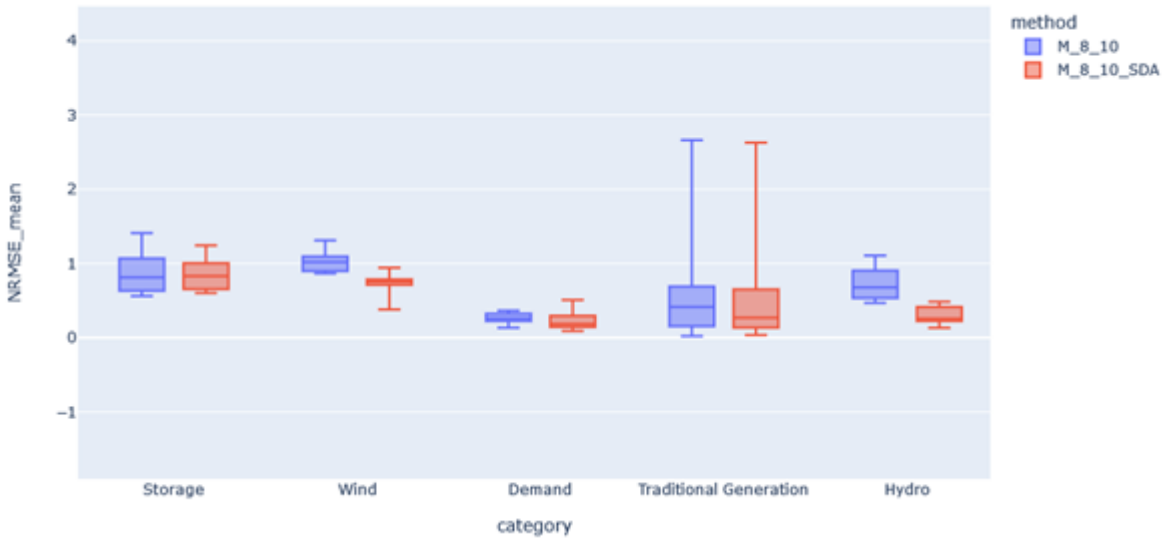


Figure 3 Box plots of the normalised RMSE per asset type for Mid 8-in-10 with and without SDA

However, as discussed above, the extreme outliers have been excluded from these plots and it was noted that some large errors are introduced by SDAs are not shown here but are a potentially significant issue introduced with this method.

[10] Box plots show the distribution and range of error data. The middle line in the box is the median, the 50th percentile, and the upper and lower edges are the 75th and 25th percentiles. The whiskers extend from 25th and 75th percentiles the to the minimum and maximum values respectively.

[11] For interpretability, the error data has been aggregated by time period prior to normalisation.

Figure 4 plots the same error metric for PV only. This graphs shows a significant increase in the range of errors when an SDA is applied. However, it should be noted that the errors presented here have been calculated for all of the historic data, therefore the seasonality and time of day characteristics of PV outputs may be contributing to this increase in error. As PV output is much lower in the UK in Autumn and Winter, the relative error would become larger at those times as the measured value is small. Furthermore, the SDAs applied in the current historic methods do not have any limits imposed on them, so for example they cannot accommodate factors such as the sunsetting, therefore scaling the baseline to match PV output at the start of a flexibility event during the day would reduce the error between the baseline and the measured data, but if the event window extends into the late afternoon that same scaling would still apply and therefore introduce an error. The simple SDA is unable to capture these characteristics relevant to PV generation.

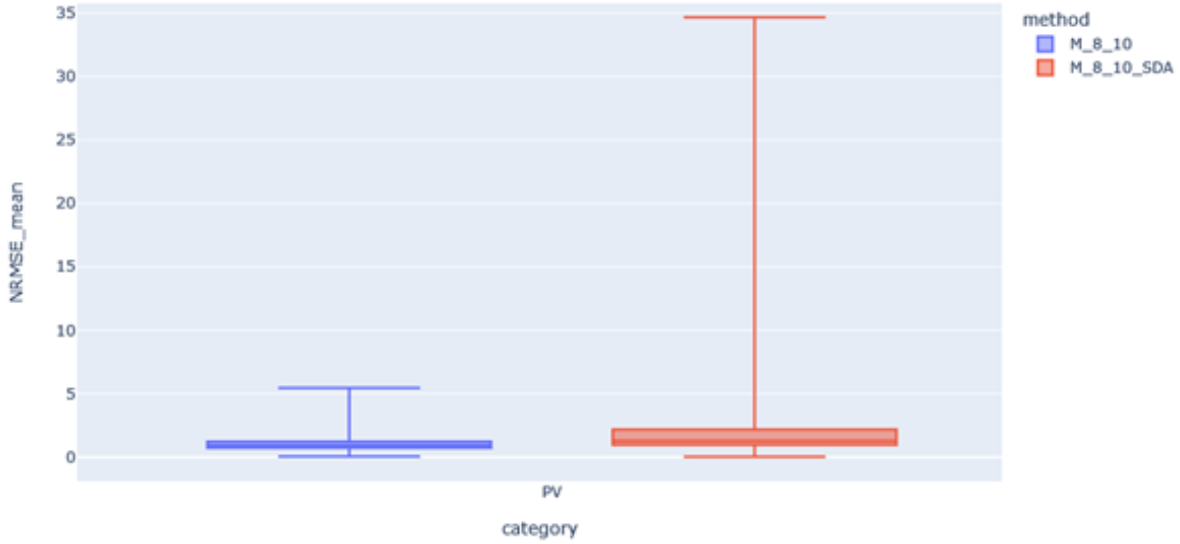


Figure 4 Box plots of the normalised RMSE for PV for Mid 8-in-10 with and without SDA

R squared values have been calculated to consider how well each of the methods is able to capture the variability of the data; a value of 1 indicates the trend is captured fully, and any score lower than zero shows a failure to capture the variability. Figure 5 plots the results for each asset type, with and without SDA. Again, for readability, extremely large errors have been excluded from the results plotted.

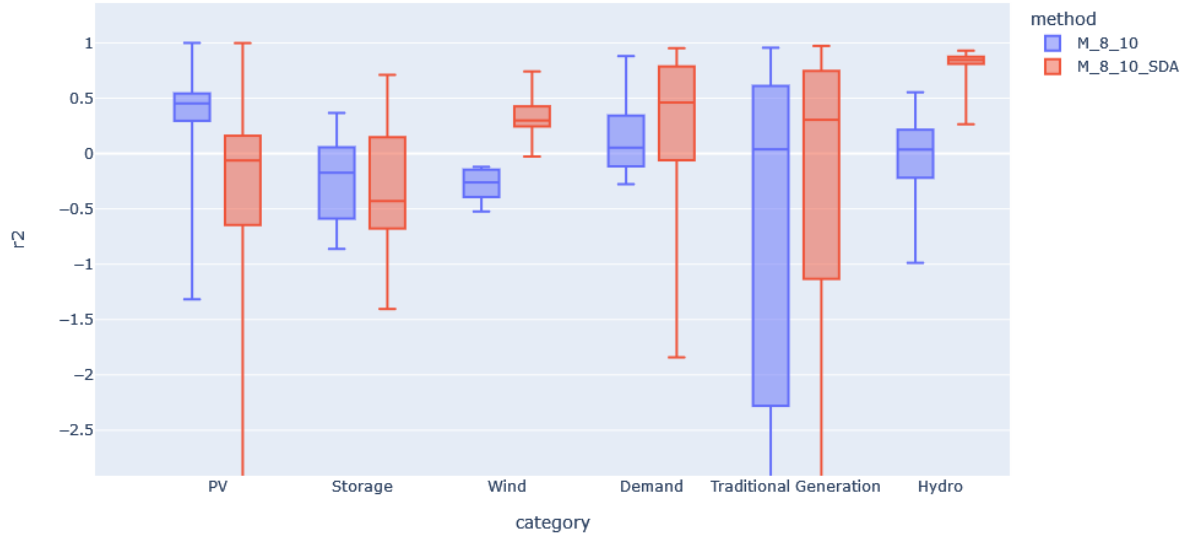


Figure 5 R Squared scores for all simulated flexibility events per asset type, for Mid-8 in-10 with and without SDA

Methods with SDA seem to be better at following the trend of data than the non-SDA methods, even though SDA methods have a worse RMS metric. This is true for most asset types, excluding PV and Storage and Demand.

Another aspect to consider is whether the methods have a tendency to either under or overestimate the performance of an asset. One way of presenting this is to count the number of times a given method produces a baseline estimate that is either higher or lower than the measured data. Figure 6 captures this data for each asset type; a negative count shows a tendency to underpredict, whereas a positive count shows a higher number of overpredictions. Note that this does not capture the magnitude of the under/over prediction, but reflects the overall susceptibility of the baseline estimate to predict either way.

Figure 6 Tendency of each method to under or overpredict

As the graph shows, the tendency to under or overpredict varies considerably by method and asset type.

In most cases the choice of method has a strong effect on consistency and accuracy of results. Overall, the historic methods seem to perform best for demand and hydro, but results for PV can be poor.

4.2 Detailed Examples for Individual Assets

This subsection provides illustrative results for single assets. The results presented in this section explore baseline estimates that have been calculated for the historic asset data with no missing data and where no flexibility services were provided.

4.2.1 Industrial Demand at 33kV

The results presented here are for a single, large scale industrial demand asset. Figure 7 plots the baseline estimate against the actual measured data for each of the timestamps within the flexibility windows simulated, for the 8 in 10 method with (bottom) and without (top) an SDA applied.

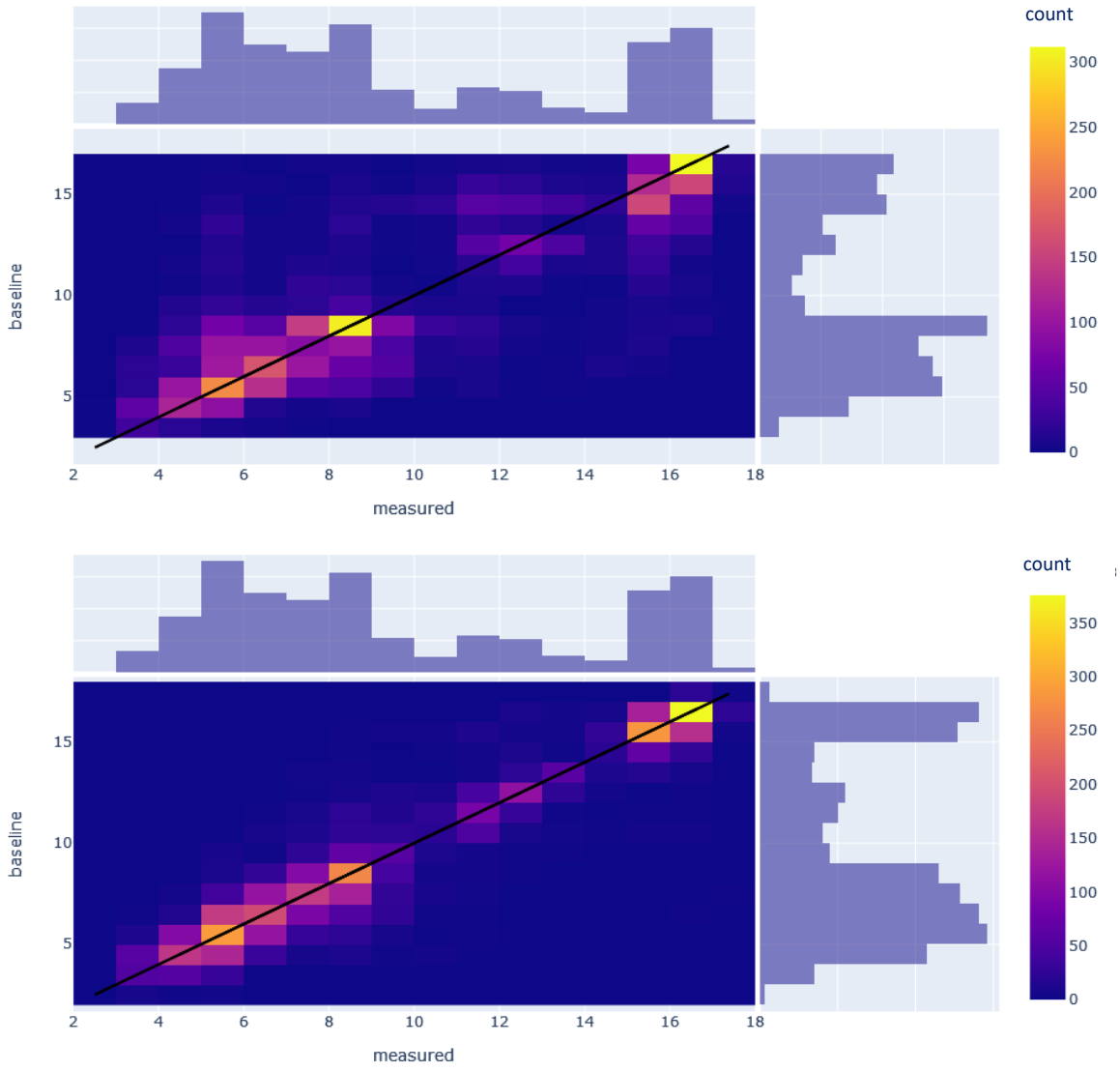


Figure 7 Joint plots of the baseline estimate vs actual measured data for an industrial demand asset with (bottom) and without (top) SDA

There are three elements to this plot. The first is the body of the graph, that plots the baseline estimate against the measured data. If the baseline estimate closely matched the actual measured data, the graphs would show A "good fit" would show the points clustered on the diagonal. Looking at the results for Mid 8-in-10 in the top plot, the points broadly follow the diagonal but there is some spread showing instances where the baseline estimate deviates from the measured data (has an error).

Comparing this to the bottom plot for Mid 8-in-10 with SDA, the spread around the diagonal is reduced, showing that SDA has been able to reduce some of the errors.

The second element of the graph is the heatmap. As there is a large amount of data plotted on these graphs, a heatmap has been used rather than plotting individual points as this would make the graph cluttered and difficult to read. The heatmap shows the density of the plots in each segment of the graph – darker colours indicates a lower density, therefore fewer results in those areas. Comparing the plots with and without SDA we can see that the density of the points around the diagonal has increased for the SDA method, showing an overall improvement in the baseline estimate.

The final element of the graph is the histogram plotted on each axis. This shows the distribution of each of the data sets. If the baseline estimate closely matched the measured data then the histograms would follow the same shape. This histogram is helpful to identify any value ranges within the measured data that the baseline estimate is either particularly good or particularly bad at capturing.

Looking at the distributions on this plot, there is there is a significantly lower density of measurements between the 10 and 15 units of power/energy in the histogram of the measured data, meaning that these values are not often observed in practice. Looking at the histogram of the baseline estimates it shows a higher density of values in this range than seen in the measured data, therefore the baseline estimate over estimates the frequency of these observations. The Mid 8-in-10 method fails to capture the absolute highest values of demand. This is likely due to the scarcity of these values, and the fact that the middle of the days is selected for the baseline calculation – it is therefore difficult to predict these using only historic days when this method is applied. In contrast, applying an SDA means that these values can be predicted, however with a tendency to slightly underpredict the value.

Considering the performance of the historic methods overall for this asset, applying an SDA increases the accuracy of the baseline estimate, shown clearly by the increase in density around the diagonal in the plot.

Looking at the distribution of the errors for each method in more detail in Figure 8, we can see that the errors are generally small, centred around zero and with no significant bias. However, some large errors are possible – the range of the error extends to +/- 10.

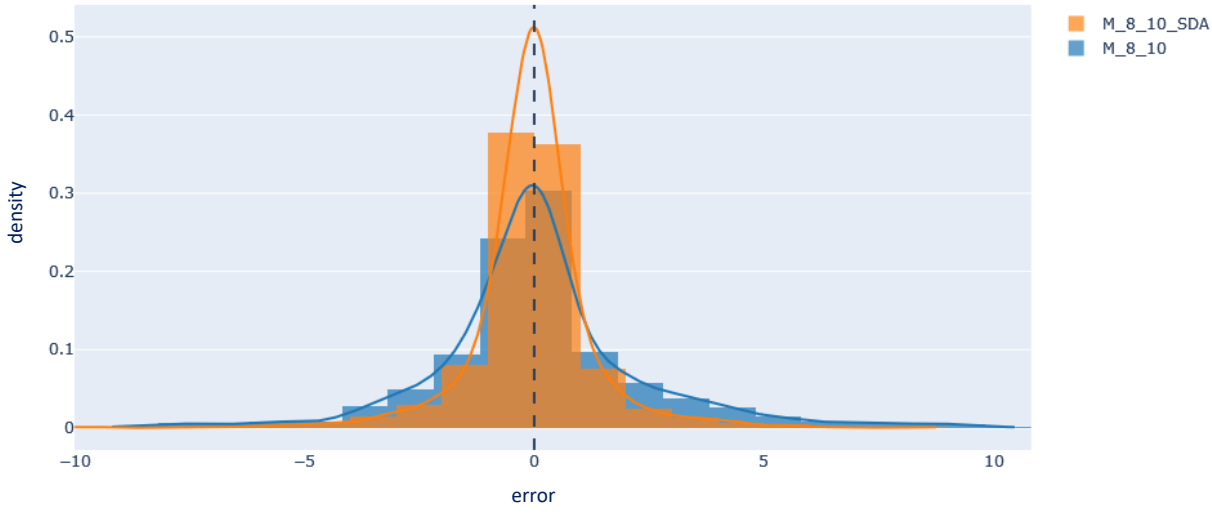


Figure 8 Distribution of errors for an industrial demand asset

It is also possible to examine how the methods perform across different seasons. Figure 9 shows that applying an SDA considerably improves the error for the Autumn season, and results for Winter are also improved.

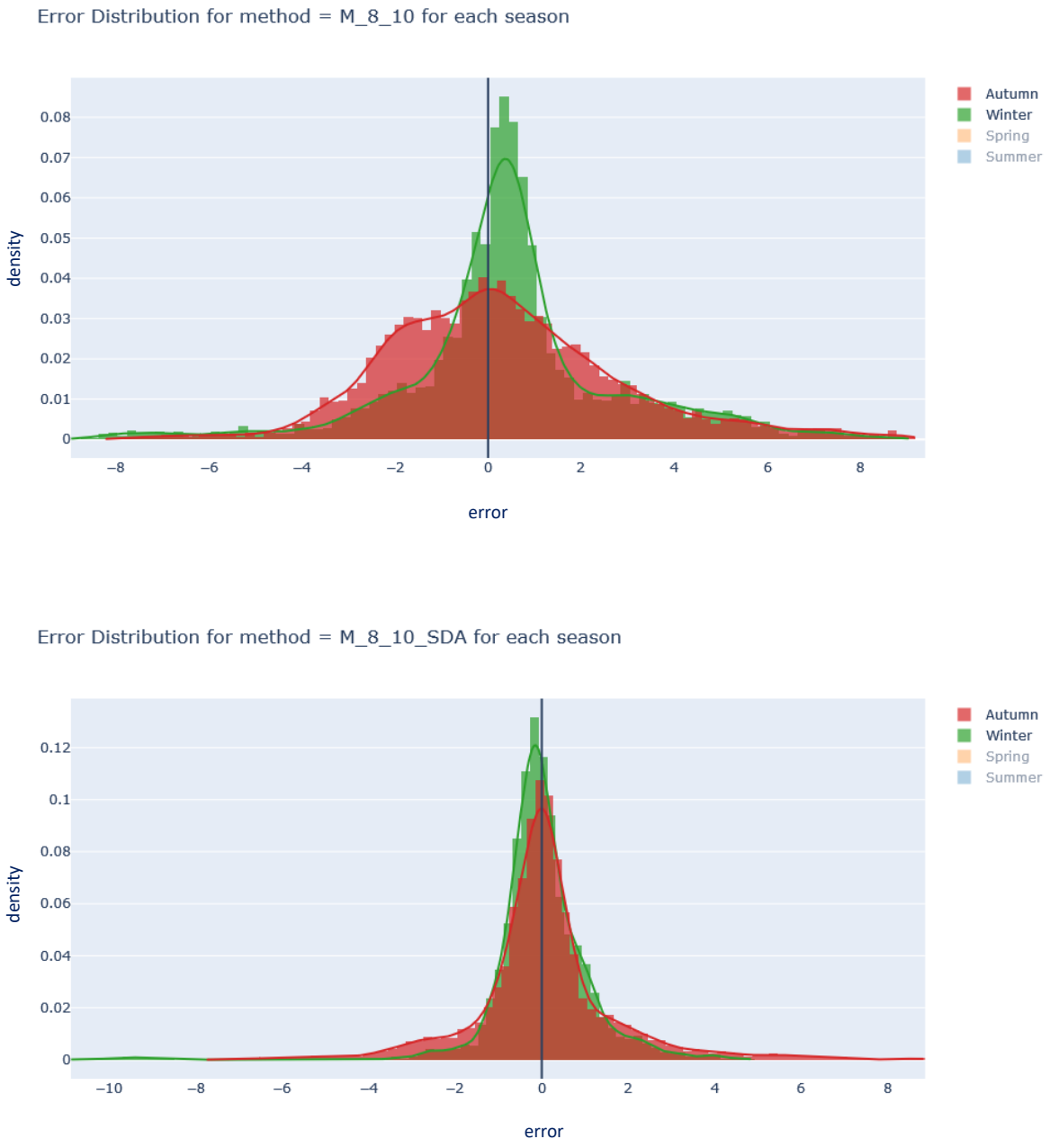


Figure 9 Distribution of baseline errors for an industrial demand asset in Autumn and Winter with (bottom) and without (top) SDA

Figure 10 shows the distribution of errors from each method in Spring and Summer. There is a small but noticeable improvement to the errors when the SDA is applied in both seasons, but with some large errors still possible.

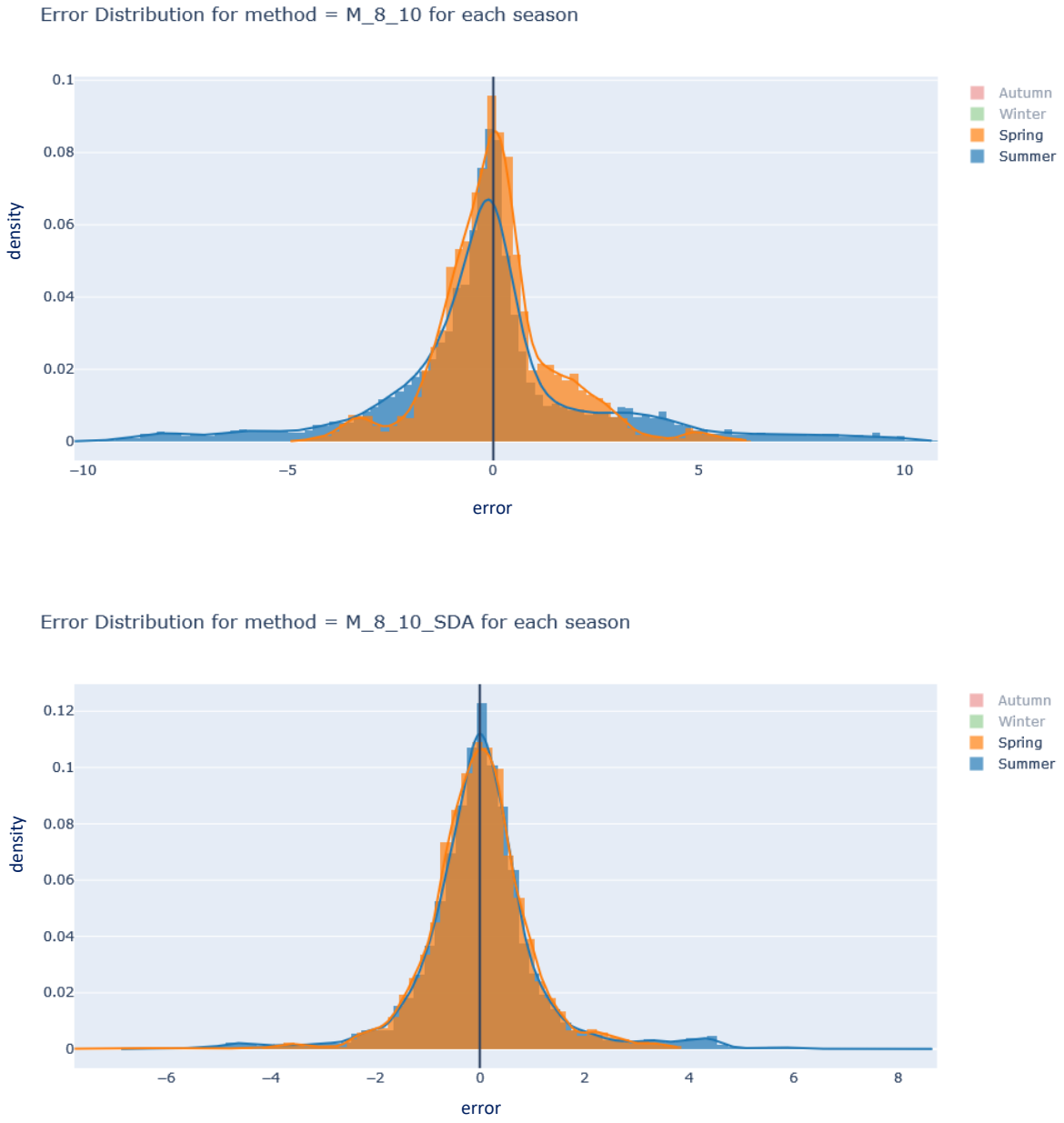


Figure 10 Distribution of baseline errors for an industrial demand asset in Spring and Summer with (bottom) and without (top) SDA

It would also be possible to consider weekdays and weekends, or other time periods of interest, for example the times when it is most likely this asset would participate in flexibility services, in order to understand if the methods perform sufficiently well during those periods. Extensions to this analysis could therefore consider the baseline errors with respect to additional contextual information.

4.2.2 PV Generation

The results presented here are for single PV asset, using the same graphical illustration described above and with the results for Mid 8-in-10 shown in the top plot, and Mid 8-in-10 with SDA shown in the bottom. However, for these plots, data for the Winter season has been removed as PV output is very low and errors during this period may obfuscate the insights that can be drawn.

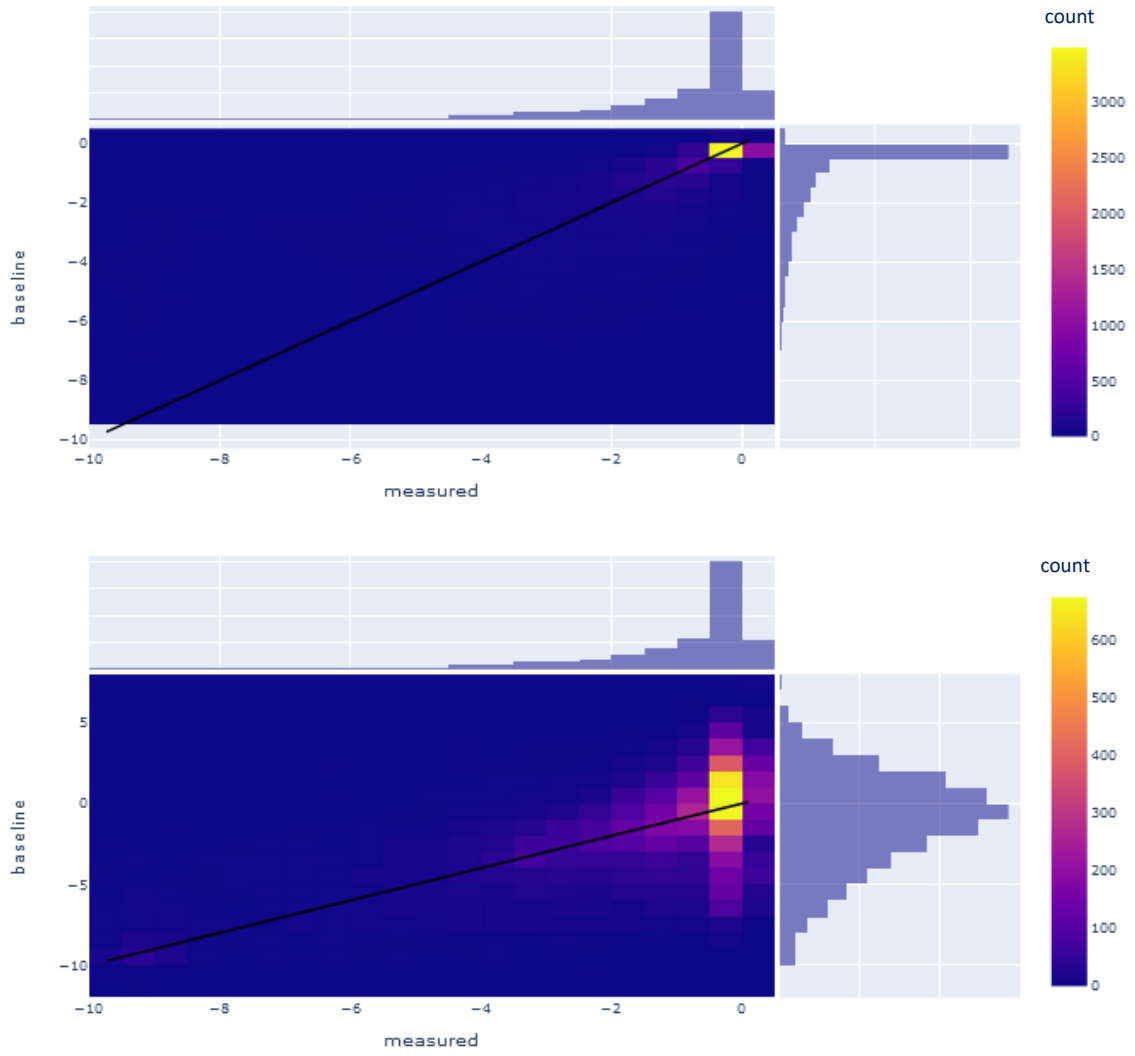


Figure 11 Joint plots of the baseline estimate vs actual measured data for a single PV asset with (bottom) and without (top) SDA

As Figure 11 shows, the introduction of an SDA increases the error of the baseline estimate. Both the heatmap and the histogram shows the inability of the Mid 8-in-10 with SDAs to follow the measured data; there is considerable spread around the diagonal, and the distribution of the baseline values deviates significantly from that of the measured data.

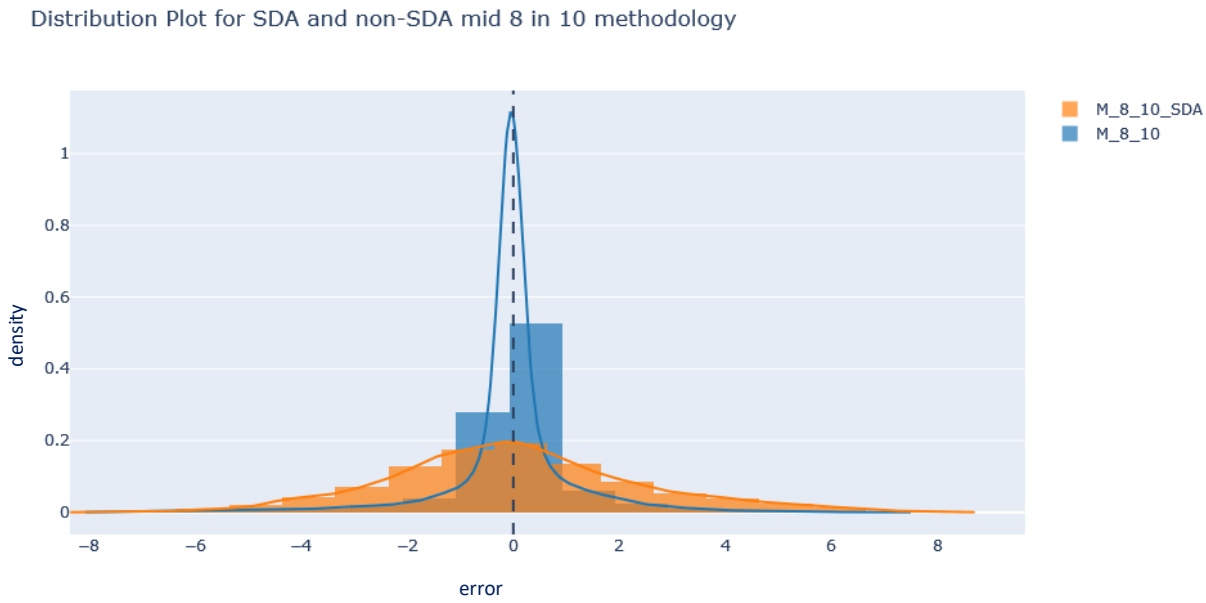


Figure 12 Distribution of errors for a PV asset

Figure 12 plots the distribution of errors for Mid 8-in-10 with and without an SDA. The distribution for Mid 8-in 10 shows a higher density around zero (an overall total number of smaller errors, but with a slight tendency to overestimate), than the Mid 8-in-10 with SDA.

However, it is important to remember that the errors presented here are for all time periods. Therefore the contextual information of when a PV asset is likely to provide flexibility, the duration of the flexibility event, as well as the volume it is able to deliver, could enhance the analysis to understand how each of the methods perform for the specific circumstances they are likely to be applied in.

4.3 Additional Analysis

The analysis was extended to consider a number of other aspects of the baselining methodologies.

4.3.1 Other Combinations of X in Y

As historic baselining methods are based on rolling averages of the metered data from days prior to the event day different combinations of days could be used, usually expressed in terms of X-in-Y. The methods examined in this report select the middle 8 of the 10 days, but other combinations have been adopted elsewhere. The results presented here consider whether any alternative combination of X-in-Y could outperform the Mid 8-in-10 method.

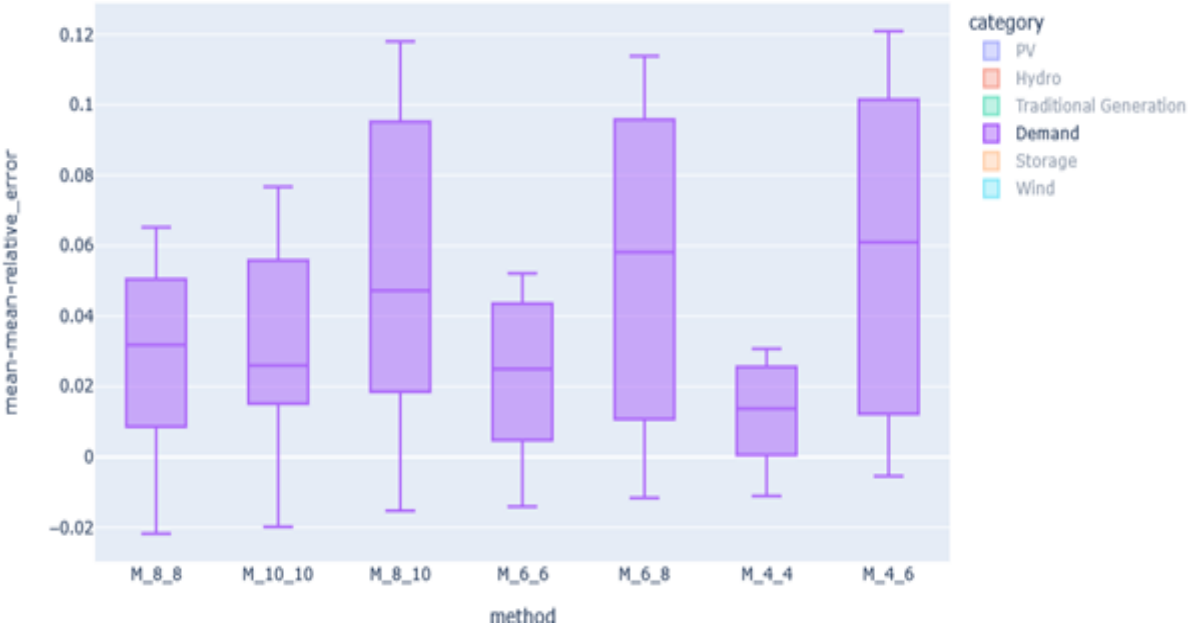


Figure 13 Variations of X-in-Y applied to demand assets

Figure 13 and Figure 14 show how different combinations of X-in-Y impact the mean relative error for demand assets and traditional generation, respectively. The results suggest that errors could be reduced by considering a smaller set of the most recent days to the event day, rather than the middle 8 of 10.

Note that as the sample size for the historic data available for each of the asset types is small these results are not conclusive. However, they are useful as they highlight the potential for other combinations of X in Y to improve the baseline estimate.

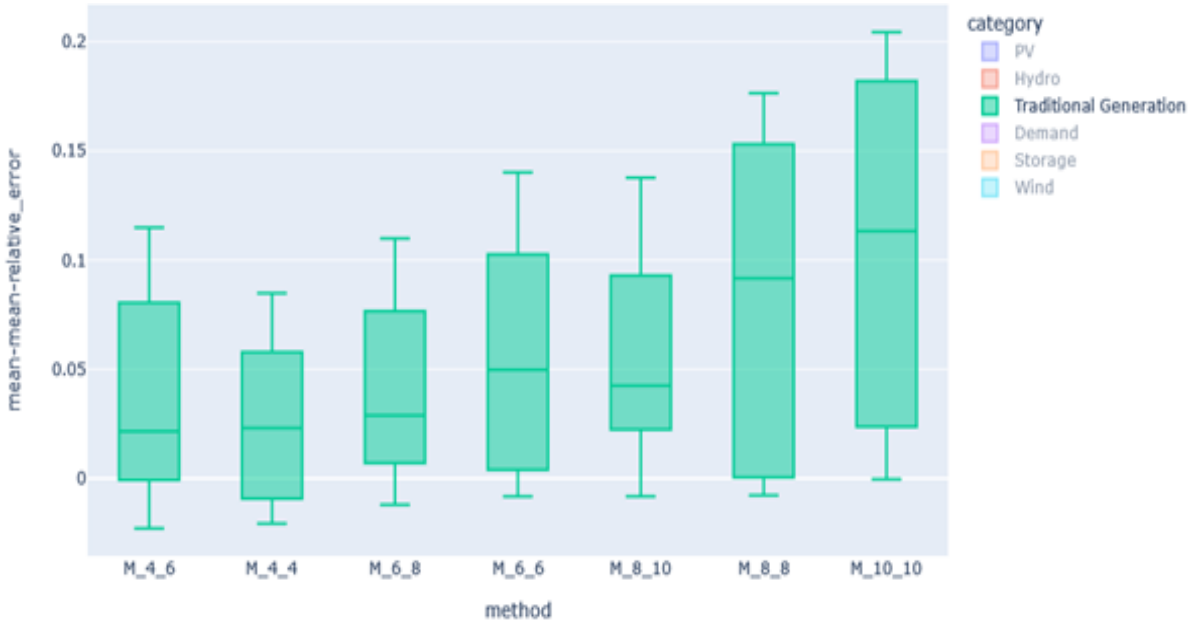


Figure 14 Variations of X-in-Y applied to traditional generation

4.3.2 Behavioural change: demand shift or demand avoidance

For some assets, particularly demand but also storage or heating/cooling, typical behaviour on a given day could be legitimately changed in order to prepare for providing flexibility. With demand, this may mean use of equipment is delayed until later in the day, for example running industrial processes or even a washing machine overnight rather than during the evening peak. This behaviour change could also be observed with storage or EV assets - if an asset is contracted to provide flexibility later in the day, it may adjust its charging pattern and charge earlier than normal in order to be able to respond to a flexibility request later on.

Flexibility from these types of assets is generally a shift in time of that demand or charging and discharging, rather than an avoidance of the behaviour entirely. It is therefore important to consider whether this behaviour change could impact the baseline estimate, and how the asset should be baselined in order to account for this. Considerations include the notification or procurement time, the length of the SDA window as well as whether or not an SDA should be applied, and the magnitude of the flexibility response required.

In order to examine the impact of this, some experimentation was conducted to shift a proportion of the demand of an asset forwards or backwards in time (for the time around the flexibility event), to see what impact this could have on the calculated baseline.

As Figure 15 shows, the methods seem to be somewhat resistant against load shifting except for when load is shifted forward into SDA window of an event. NTVV's larger event window will make it more sensitive to this as it has an SDA window of 4 hours prior to the event rather than the than the window of 2 hours used in the Mid 8-in-10 with SDA. Reducing an assets normal demand to then increase it later (shifted backwards in time) has the smallest impact on the baseline estimate.

In practice, consideration should be given to how an asset is likely to behave, as well as what is acceptable for the baselining for a given flexibility service, in order to determine which method is appropriate for baselining when behaviours are shifted rather than avoided.

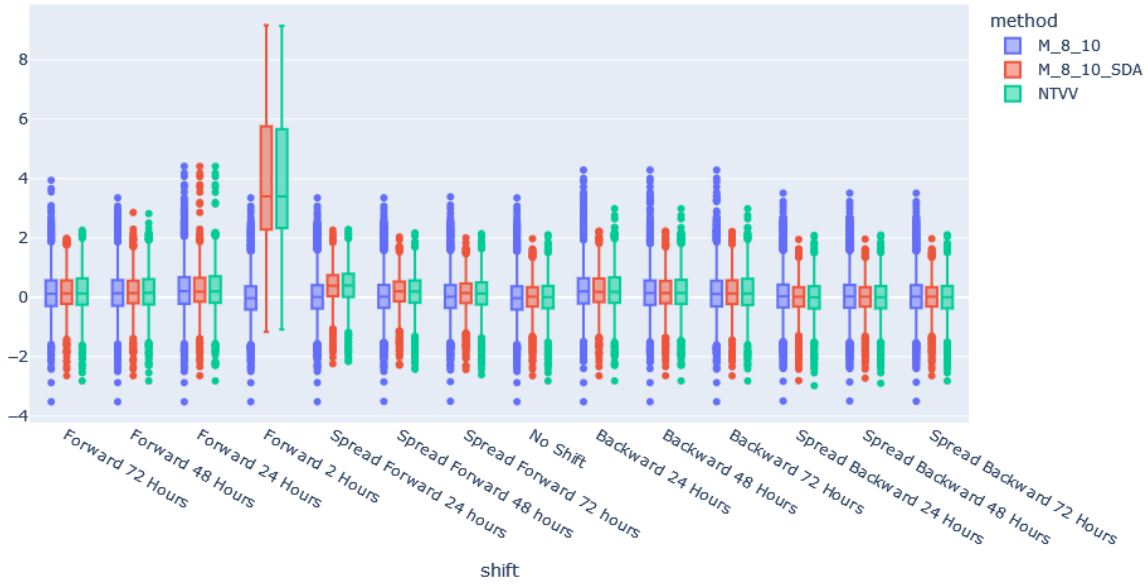


Figure 15 Experimentation with shifting behaviour patterns and the impact on the baseline error

5 Conclusions and Recommendations

Baselining methods that use historic data, either with or without same day adjustments are simple, transparent, easy to use and have modest data requirements. This makes them accessible and easy to implement.

In many cases, the error in the baseline estimate is typically small, centred around zero without significant bias. However, this is not consistent for each asset types and large errors are possible, particularly for PV generation. In most cases the choice of method has a strong effect on consistency and accuracy of results. Overall, the historic methods seem to perform best for hydro and large-scale demand, but results for PV are poor which suggests that alternative methods could be more suitable.

The analysis has also shown that the tendency of the methods to over or underestimate the measured data is not consistent across asset types. Applying an SDA to the historic method helps to minimise this tendency for most results for some types of assets, but it also introduces severe under and/or over prediction in edge cases. Further consideration should be given to time and product type a generator may be offering in order to determine whether the methods perform sufficiently well for the periods when the asset is likely to provide flexibility.

The results presented in this analysis cover a range of asset types and historic years, however the sample size for some asset types is very small. Therefore, the conclusions and recommendations presented in this section are indicative of the method performance but may not be conclusive for those assets where only a few examples are available e.g. hydro, demand, wind and storage. Therefore, further investigation of historic method performance is recommended, using a wider range of asset data across each of the available asset types. Further considerations and recommendations are discussed below.

Application of same day adjustments

The application of same day adjustments has been shown to reduce the baseline error in some cases, however it can also increase errors and produce baseline estimate values that are not possible in practice, for example negative values for generation. It is therefore important to consider that while applying an SDA does help to match baseline to the data on the event day, the fact that there are no limits associated with adjustment (shift, or scaling) means that that there is the potential to introduce significant errors. It is also possible that applying an SDA fails to capture other characteristics of the asset during or at the end of the flexibility event, for example sunset and PV output – the same adjustment made to a baseline to match the output before the flexibility event starts earlier in the day could introduce an error later on.

The impact on the baselining error is not the only consideration when applying SDAs. It is possible that manipulation of the baseline could happen, therefore the timings of the notification period and event window should be considered. For example, if the SDA window overlaps with strategic pre-emptive behaviour from by the asset, the baseline may be exaggerated. This could then impact apparent volume of flexibility delivered during the flexibility event. On the other hand, an asset may be able to manipulate their behaviour in this way deliberately. Careful consideration should be given to when SDAs should be applied, and any susceptibility to manipulation.

Opportunities for expanding the performance assessment analysis

There are further questions about their performance that could be examined as part of the ongoing development of baselining methods as they are used throughout the TRANSITION trial periods. An extended analysis would also benefit from a larger number of datasets across different asset types, to

determine whether conclusive decisions could be made for which available method performs best for each asset type. This would also enable a more comprehensive examination of the impact of data quality on baselining estimates to inform any decisions on minimum data requirements.

Similar analysis to what has been presented in this report could be expanded to include the possible volume of flexibility that an asset is able to provide, for various products and services or times of day and year, in order to understand the impact of the error in the baselining estimate relative to the magnitude of flexibility that could be delivered by the asset, and at those times an asset is likely to provide flexibility. This could quantify the magnitude of the error relative to the actual flexibility response of the asset, and then any practical implications for delivery verification and payment could be considered in more detail.

Consideration should also be given to portfolios of assets providing flexibility collectively, in order to understand how the errors of the individual assets behave when aggregated over several providers and asset types. Aggregated portfolios of assets are already beginning to provide flexibility collectively, and will likely be the main providers of flexibility in distribution level flexibility markets, so it is important to understand the impacts of errors for these types of providers. An extension of this could also include an examination of the differing flexibility service values and payment structures, to understand if there is any significant impact on service delivery and therefore payment, and how this may vary depending on the payment structures in place.

If other methodologies are developed or adopted, for example regression-based methods, it would be beneficial to conduct a comparative analysis of their performance with respect to the existing historic baselining methods. This analysis should examine whether any new methods can consistently reduce the baseline error, and to understand which circumstances, flexibility produces, or assets may be best served by the available methods.

Development of regression-based methods

The analysis presented here has shown that there are opportunities to improve upon the performance of historic data-based methods for baselining. Development of additional methods, for example regression-based methods, are more complex and data intensive than the more simple historic methods, but could produce more accurate baseline estimates, especially for those assets types such as PV that are not very well served by historic methods. It is important that baselining methods are developed with simplicity, transparency, and replicability in mind, but also inclusivity – and regression-based methods could improve inclusivity, in particular for PV and other weather driven renewables or highly variable assets.

Regression-based methods are models that use additional and external data (data that is not simply historic measured output data) to model the behaviour of an asset, thus determining the relationship between the external data and the output of an asset in order to calculate a baseline from that data. Data for these types of methods could include time variables such as time of day, time of year, as well as other factors such as temperature data, weather data, or even price data. The point of application would determine what type of dataset is used. For example, validating flexibility service delivery would require historic external data calculate the baseline using the actual data from the event day. Calculating a baseline ahead of service delivery would require forecast data, such as weather forecast data. The accuracy of any future calculated baseline would depend on the accuracy of those forecasts.

Several of the industry studies reviewed in Section 2.3 note that there may be barriers to developing and adopting regression-based methods, such as the availability of data or the data processing requirements. However as digital technology has developed these may no longer present a challenge.

Furthermore, as these methods are based on external data (rather than simply historic metered data) they may actually increase the inclusivity of flexibility service participation.

The discussion in ELIA's report[6] supports the development of regression based methods, and shows a comparison between adjusted and unadjusted historic baselines. Regression based methods are shown to outperform unadjusted methods, but not those which have applied SDAs. However, the analysis shown here has highlighted that this is not true for PV assets, therefore regression-based methods could be very valuable for PV and other renewable generation where the performance of historic methods is inconsistent.

In addition, the use of external data would also mean that regression-based methods would become more valuable as the frequency of flexibility events and the number of available services increases. The methods based on historic data examined in this report, and those used more widely in industry, exclude periods of time or full days of data where an asset was providing flexibility from the baselining calculation. It would not take much of an increase in the frequency of flexibility events and services for this to soon mean that searching for eligible days to use in the historic baselining method may mean that those could be weeks prior to given flexibility event, and therefore this data could reflect very different operational or weather conditions those seen on the event day. Using methods that do not solely rely on historic metered data could mitigate this issue, for example if weather data was used for renewables generation, then this issue could be mitigated.

Work on regression-based methods will be the next focus of TNEI's work with TRANSITION.[12]Following this, it would be possible to conduct a similar performance assessment, as well as a comparative analysis between the historic and regression methods. This would help to understand which methods perform better in which circumstances, or for which asset types.

Further consideration of the practical implications of baselining estimates

Future work should consider practical implications of applying a given baselining method, including:

- If using a method that applies an SDA, what combination of notification period for the flexibility event, and adjustment window for the SDA is appropriate or acceptable?
- What are the practical implications of errors in the baselining estimates for the flexibility service valuation and settlement procedure?
- What level of error is acceptable? What further contextual information should be considered to determine whether the error is acceptable? This could include the likely volume of flexibility an asset can deliver, the times it is likely to provide flexibility, as well as the flexibility product and constraint type.
- What methods should be adopted if the frequency of activation of the flexibility service or product increases?
- What if an asset is delivering several different flexibility services, how is this captured and reflected in the data used for calculating the baseline?

[12] The specific focus will be on regression baselines for PV. This is due to both practical issues around data availability, as well as materiality (based on the performance analysis presented here can, we see that PV is not well served by the historical baselining methods).

There are also significant considerations here for how baseline estimation methods should develop to meet the developing needs of flexibility service markets as they grow in frequency and type of product, and as they continue to roll out across GB at all network levels.

Encourage wider use of the historic methods in the Flexibility Baselining tool

As the use of historic baseline methods is continuing throughout the TRANSITION project trial periods, experience with the methods will grow. The variety of asset type and dataset will increase, therefore increasing the knowledge of how these methods perform, how they are used, as well as any potential concerns.

The methods are also available for industry sector and public use through the Flexibility Baselining Tool[1] and are being rolled out for DNO flexibility services across GB through the ENA Open Networks programme. The Tool itself is openly published, as are the mathematical specifications of the historic baselining methods. Publishing the tool and methods openly supports transparency of the methodologies adopted for flexibility baselining and allows for broader use and trialling of the methods than would be possible through TRANSITION alone. This enables a wider range of users to understand the performance of the historic baselining methods for their own data and broadens the range of asset types and data used in the calculations. Overall, this will help DNOs, FSPs and the wider energy sector to better understand the performance of the methods for a variety of data and asset types, and can also help to capture any potential concerns or suggestions they have for future development.